NOVEL, CONSERVED RNA SECONDARY STRUCTURES IN MHV-A59, BOVINE CORONAVIRUS (BCoV) AND MERS-CoV

An Undergraduate Research Scholars Thesis

by

VINATHI SAINAGA POLAMRAJU

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Julian Leibowitz

May 2018

Major: Biomedical Sciences

TABLE OF CONTENTS

ABSTRACT

Novel, Conserved RNA Secondary Structures in MHV-A59, BCoV and MERS-CoV

Vinathi Sainaga Polamraju
Department of Microbial Pathogenesis and Immunology
Texas A&M University

Research Advisor: Dr. Julian Leibowitz
Department of Microbial Pathogenesis and Immunology
Texas A&M University

Betacoronaviruses are a subgroup of viruses in the family Coronaviradae known to cause an array of diseases in humans and animals. In this study, we aim to determine the RNA secondary structures of Mouse Hepatitis Virus, strain A59 (MHV-A59), the best studied betacoronavirus, and closely related betacoronaviruses, BCoV and MERS-CoV to identify novel, conserved secondary structures within their genomes. To accomplish this, we infected DBT, HRT, and Vero-E6 cell cultures with their respective virus stocks: MHV-A59, BCoV, and MERS-CoV. Upon viral clarification and titration, we obtained virus titers between 1.0 and $1.42x10^7$ pfu/mL and purified viruses via differential and sucrose density gradient centrifugation. Subsequently, we extracted the viral RNA and reacted it with SHAPE-MaP reagent 1-methyl-7-nitroisatoic anhydride (IM7) which probes for and forms adducts with conformationally flexible ribose 2'-hydroxyl groups in the RNA. The derivatized RNA is reverse transcribed in the presence of Mn++ causing misincorporation at adduct sites. This induces mutations in the cDNA transcripts which are incorporated into a cDNA library. Thus, deep sequencing of this cDNA library provided us with an avenue to create relatively accurate RNA secondary structure models using Shannon entropy and pairing probability models. High-confidence regions, characterized

by low Shannon entropy and low SHAPE reactivity, were selectivity visualized. The folding models generated by FORNA were visually analyzed for conserved structures and covariation. Three conserved secondary structure models, located in open reading frame (ORF) 1b, were isolated and are thought to be important in translation and could serve as binding sites for host or viral proteins. Further studies will include conducting site-directed mutagenesis to understand the functional role of these secondary structure models and utilizing ShapeKnots analysis to probe for pseudoknots.

Keywords: **Betacoronaviruses, MHV-A59, BCoV, MERS-CoV, SHAPE-MaP, Shannon entropy, SHAPE reactivity**

DEDICATION

To my family and friends, I could not have done this without you. Thank you for your constant encouragement and support.

ACKNOWLEDGEMENTS

I would like to thank my faculty advisor, Dr. Julian Leibowitz, for his guidance and motivation throughout the course of this research. Without his passion and perseverance, this project would not have been possible. I would also like to extend my gratitude to the members of the Leibowitz Lab, Mr. Drew Nunn and Dr. Joseph Tingling, for their support and assistance.

Thanks also go to my friends and colleagues and the department faculty and staff for making my undergraduate experience at Texas A&M University memorable. Additionally, I would like to express my acknowledgements to the Weeks Lab and the Schaniuk Lab for their support in RNA visualization. As well as the AgriLife Genomics and Bioinformatics Service and Texas A&M Institute for Genome Sciences and Society for their help with Illumina sequencing. I would also like to extend my gratitude to Dr. Byung-Jun Yoon from the Department of Electrical and Computer Engineering at Texas A&M University for his assistance in bioinformatic analysis and Dr. Chien-Te K. Tseng from the University of Texas Medical Branch.

Finally, thanks to my mother and father, Varalakshmi and Srinivasa Polamraju, for their encouragement and strength and my sister, Tanmayee Polamraju, for her inspiration and motivation.

ABBREVIATIONS

MHV-A59	Mouse Hepatitis Virus (strain A59)
BCoV	Bovine Coronavirus
MERS	Middle East Respiratory Syndrome
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
SHAPE-MaP	Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling
1M7	1-methyl-7-nitroisatoic anhydride
UTR	Untranslated region
TRS	Transcription regulatory sequence
MSE	MOPS Saline EDTA

CHAPTER I

INTRODUCTION

Coronaviruses

The Nidovirus superfamily encompasses a family of viruses known as coronaviruses that are divided into four genera based on antigenic reactivity, later confirmed by genomic sequencing. These four groups are known as alpha, beta, gamma, and deltacoronaviruses.[4,6,21] Betacoronaviruses, in particular, are further subdivided into four lineages: a, b, c, and d.[6,21] Moreover, coronaviruses are capable of inflicting disease in a wide variety of animals and humans. The most widely studied coronavirus, mouse hepatitis virus (MHV), belongs to the betacoronavirus genus and is recognized as a model system for studying various central nervous system (CNS) diseases, including encephalitis, multiple sclerosis, and acute hepatitis.[4] Bovine coronavirus (BCoV), another member of the betacoronavirus genera, causes respiratory diseases in cattle and continues to be a problem for beef and dairy industries.[4] On the other hand, coronavirus-induced infections in humans normally amount to the common cold. However, recently emerging coronaviruses, such as severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), have reaped devastating effects worldwide. SARS-CoV is responsible for the severe acute respiratory syndrome (SARS) outbreak of 2002 in China.[4] According to the World Health Organization (WHO), SARS affected over 8,000 people with a 10% mortality rate and spread to over two dozen countries.[4,30] MERS-CoV caused the Middle East respiratory syndrome (MERS) outbreak of 2012 in Saudi Arabia.[4] According to the WHO, MERS affected over 1,700 people with a 37% mortality rate.[4,30] This thesis will focus on three betacoronaviruses, namely MHV, BCoV, and MERS-CoV.

Phylogeny

Coronaviruses are members of the coronavirinae subfamily within the coronaviridiae family of the Nidovirus superfamily.[37] As mentioned previously, coronaviruses are divided into four genera (alpha, beta, gamma, and delta) within which the betacoronavirus genus is further classified into specific lineages (a, b, c, and d).[4,6,21] The following section will provide a brief overview of the characteristic viruses belonging to the various lineages.

Mouse Hepatitis Virus (MHV)

Mouse Hepatitis Virus (MHV), a member of lineage a, is considered a model organism for studying various hepatic, neurologic and enteric infections.[35] The various strains of MHV are capable of utilizing the same host cell receptors to gain access to multiple organs. The most widely studied strains of MHV are neurotropic in nature and include JHM and A59.[35] Both are responsible for demyelinating encephalomyelitis, the human equivalent of which is multiple sclerosis (MS). Consequently, during clearance, myelin destruction ensues resulting in fatal acute encephalitis in JHM infected mice.[35] Unlike JHM; however, MHV-A59 is also capable of infecting the liver.[35]

Bovine Coronavirus (BCoV)

Bovine Coronavirus (BCoV), another member of lineage a, causes infections of both the upper and lower respiratory tracts as well as the intestines. This virus infects housed, adult cattle with diarrhea, more commonly known as Winter Dysentery, worldwide.[18] Coronavirus OC43, which causes the common cold, has been recognized as the human counterpart of BCoV.[18] Variants of BCoV are also capable of infecting dogs with respiratory infections and humans with diarrhea.[18]

Middle East Respiratory Syndrome Coronavirus (MERS-CoV)

Middle East Respiratory Syndrome Coronavirus (MERS-CoV), a member of betacoronavirus lineage c, is a recently emerged cause of fatal respiratory illness in humans.[23] MERS-CoV causes Middle East Respiratory Syndrome (MERS) in humans and was first reported in Saudi Arabia in 2012.[10,23] Since then, it has been reported in 27 other countries including the United States, North Africa, and Europe.[23] Although MERS-CoV is inefficiently transmitted amongst humans, MERS-CoV infection carries an approximate 35% mortality rate.[23] Human-to-human transmission of MERS-CoV is generally limited to unprotected, direct human-human contact in health care settings. While formal proof of the exact route of transmission is lacking, dromedary camels are suspected of being the major reservoir host.[9,23] Symptoms of MERS include respiratory distress, such as cough and dyspnea, gastrointestinal complications, such as diarrhea, and renal failure.[23] Due to its recency, there is no current vaccine available for MERS-CoV infections.

Genome Organization and Replication

Genome Organization

Coronaviruses are characterized by a single-strand, positive sense RNA with genome sizes ranging from 27-32 kb.[3] Amongst all coronaviruses, the 5' two-thirds of the genome encodes a replicase locus and the 3' one-third encodes various structural proteins and accessory proteins not required for in vitro growth. Two overlapping large open reading frames, which extend from about nucleotide 210 to encompass two-thirds of the genome, encode the proteins which make up the replicase complex.[12] This region is translated as two large polyproteins, orf 1a and orf 1ab, that can be further co-translationally cleaved into 16 proteins.[12] Ribosomal frameshifting from orf 1a to orf 1b, which utilizes a slippery sequence and an RNA pseudoknot, is required for the expression of the two aforementioned polyproteins. These proteins include

proteases, RNA modification enzymes, polymerases and helicases.[3] Also located at the 5' end is an approximately 75 nucleotide leader sequence, which contains a transcription regulatory sequence (TRS) at its 3' end, and an untranslated region (UTR) that contains bulged stem loops important for viral transcription and replication.[3,5] Transcription regulatory sequences (TRS), positioned at the beginning of both structural and nonstructural genes, serve as binding sites for RNA polymerase and are important in orchestrating genomic expression.[5] The structural proteins of the 3' end are arranged in the following order from the 5' end to the 3' end: hemagglutinin esterase (HE), spike (S), small membrane (E), membrane (M), nucleocapsid (N), and internal protein (I) located within the N gene.[12] More specifically, the HE protein is found only in betacoronaviruses. The crown-like morphology characteristic of coronaviruses is due to the spike (S) protein, found as a homotrimer and the HE protein, if present.[3,12] The M and E protein are additional transmembrane proteins important in virus assembly.[27] The helical capsid structure found within the envelope is formed by the nucleocapsid protein as it complexes with the RNA genome.[3] Similar to the 5' end, the 3' end contains an untranslated region (UTR), approximately 300-500 nts, that is composed of a bulged stem-loop, a pseudoknot, a hypervariable region, and a poly-A tail depending on the virus.[12] However, since the stem loop and pseudoknot regions overlap, they cannot be formed simultaneously. Thus, the different structures proposed are thought to be important for controlling alternate stages of viral RNA synthesis.[5] Figure 1 is a representation of the genomic arrangement of MERS-CoV for reference:

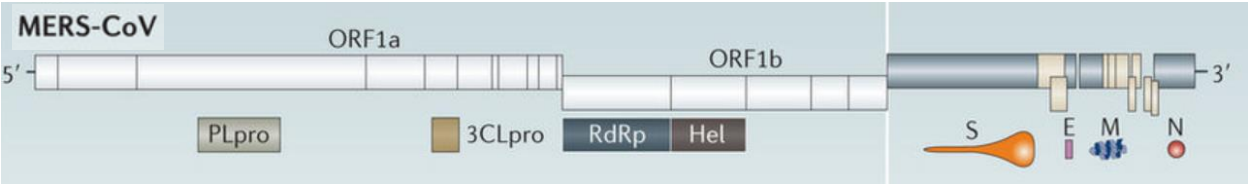

Figure 1. Arrangement of MERS-CoV genome. From Zumla *et al*., 2016.

Coronaviruses can be further distinguished by the presence of interspersed, accessory nonstructural genes that are not vital for replication.[3] These proteins differ in sequence, number, and function amongst coronavirus groups.

Replication

Coronaviruses use the spike (S) protein and the HE protein, if present, to attach to cell surface molecules.[27] The single-strand, positive sense RNA is then deposited into the host cell, which marks the beginning of the replication process.[29,34] Sub-genomic and genomic mRNAs are produced via negative sense intermediates during viral mRNA synthesis.[29,34] This RNA genome undergoes translation to produce viral protein products, including RNA-dependent RNA polymerases (RdRp) that pauses at TRS sequences to either continue RNA synthesis to the next TRS or transcribe the leader sequence located at the 5' end.[29,34] Characteristically, during the synthesis of these sub-genomic mRNAs, the leader and body TRS segments fuse together allowing for the elongation of negative-sense RNA.[29,34] The smaller sub-genomic positive mRNA strands, used to produce the structural proteins that form the capsid, and new positive sense RNA genomes are produced by the negative-sense RNA intermediates.[29,34] After the N protein binds to the newly synthesized RNA genome, the M protein becomes embedded into the membrane in the endoplasmic reticulum along with the S and HE proteins.[27] During virus budding, mediated by the M protein, the fully formed virus particles are exocytosed into the extracellular space via Golgi bodies.[27] Figure 2 depicts the morphology of SARS-CoV for reference.

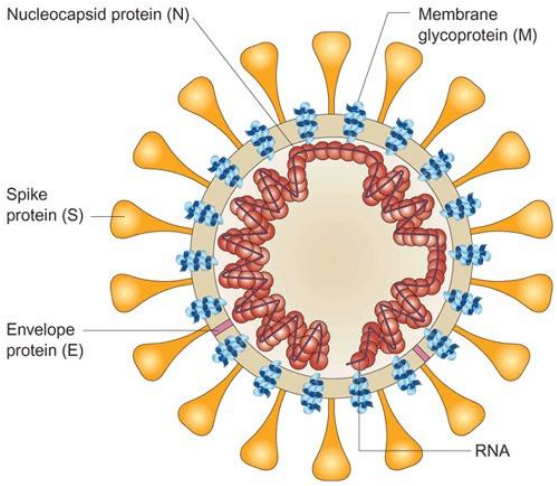

Figure 2. SARS-CoV morphology. From Nicholls *et al*., 2008.

Morphology

Historically, coronavirus identification depended solely on their characteristic morphology; however, recent biochemical and serological profiles have become available. Measurement analysis using negative staining has revealed total diameters ranging from 75 to 160 nm.[14] Coronaviruses can be generally characterized by their spherical shape and widely-spaced surface projections.[14,28] The surface projections can take the following forms: the typical bulbous, "tear-drop", or rod-shaped with a T-shaped structure at the distal end.[14,28] Moreover, these projections vary in length, ranging from 12 to 24 nm.[14,28] Some coronaviruses, including IBV and hemagglutinating encephalitis virus, can have different surface projection structures; however, others have solely one type. These surface projections are composed of similar glycopolypeptides that are arranged differently, resulting in the slightly different morphology.[8,28] Regardless, the three different surface projection structures are due to the S protein and share the same biological functions, recognizing receptors on the target host cell and viral entry into the cytoplasm.[8,14]

RNA Secondary Structure

Role of RNA Secondary Structure in the 3' and 5' Untranslated Regions (UTRs)

Complementary regions on a single RNA molecule can create double helical stretches with interspersed loops, also known as the secondary structure of RNA. The linear genome is capable of folding into crucial cis-acting elements.[19] The RNA secondary structure plays a vital role in biological regulation, including altering stability and translation and transducing signals, and is hence widely studied.[19,20] Functional RNA molecules can be distinguished by their characteristic secondary structure, an essential precondition for their function.[19,20] Thus, through the course of evolution, many RNA secondary structures have been highly conserved.

The RNA secondary structures of the 5' and 3' UTRs of coronaviruses have been widely credited for providing stability and participating in inter- and intra-molecular interactions.[2] Specifically, these include interactions between cellular and viral proteins during translation and replication and other RNA-RNA interactions.[2] Moreover, the cis-acting sequences of betacoronaviruses, including MHV, BCoV, and MERS-CoV, display remarkably similar secondary structures despite their divergent genomic sequences.[14] The 5' UTR is characterized by unique stem loops (SLs) that are numbered in order from the first nucleotide base. MHV, BCoV, and MERS-CoV share conserved structures for SL1, SL2, SL4, and SL5ABC.[8,13,20] Perhaps one of the most distinguishing elements of the RNA secondary structure for the three aforementioned viruses is the folding of the most distal end of the 5' UTR, the sixth and seventh stem loops, namely SL6 and SL7.[8,13,20] In MHV, two separate stem loops are formed; in BCoV, a forked stem loop is formed; in MERS-CoV, two bulged stem loops are predicted.[8,13,20] Moreover, an additional stem-loop (SL3) has been recognized in BCoV that participates in configuring the leader TRS sequence into a hairpin loop.[13] This specific structure is important in the replication

and transcription of BCoV.[13] While similar structures have been revealed in MHV, they are likely not stable.[13] Regardless, the relatively conserved nature of the 5' UTR of the three viruses further underlines the importance of this region in viral RNA synthesis and replication.

Similar to the 5' UTR, the 3' UTR contains cis-acting elements important in viral replication. The poly-A tail portion of the 3' UTR has also been noted for its influence in initiating replication and minus-strand RNA synthesis in MHV.[13,19] A bulged stem-loop, located at the most 5' end of the 3' UTR, is thought to be conserved amongst MHV, BCoV, and MERS-CoV.[8,13,20] Just downstream of the bugled stem-loop is a hairpin stem-loop that can interconvert into a hairpin-type pseudoknot.[8,13,20] A pseudoknot is a unique secondary structure which consists of two stem-loop structures where half of the first stem-loop is intercalated between the second stem-loop. The pseudoknot structure in the 3' UTR is also conserved amongst the three viruses.[8,13,20] However, the primary nucleotide sequence of this region is only partially conserved, suggesting that this structure plays an important regulatory function.[8,13,20] Moreover, studies show that the bugled stem loop and the neighboring pseudoknot overlap and cannot be formed simultaneously.[8,13,20] Thus, it is hypothesized that these structures regulate the transition occurring during viral RNA synthesis.[8,13,20] The following section will focus on various methods used to predict these RNA secondary structures.

Methodologies for RNA Secondary Structure Visualization

The secondary structure of RNA is defined by intramolecular interactions, or pairings, of complementary sequences of at least two base pairs. Bases can pair in a canonical (A-U, G-C, etc.) or noncanonical (G-U, A-G, etc.) fashion.[26]

Comparative Analysis

One of the earliest approaches in the field of RNA secondary structure prediction is comparative analysis. This approach allows one to infer the secondary nature of RNA using phylogenetic comparisons.[26,31] The underlying principle is that interacting base pairs are conserved in multiple homologous RNA sequences.[26,31] The term homologous refers to the fact that the sequences share a common ancestor and are predicted to have similar higher order structures. Moreover, the conserved pairings can also include base pair compensations. These refer to evolutionarily conserved structures that surprisingly display diverging sequences.[15] For example, a G-C canonical base pair may be substituted for an A-U base pair in another sequence. Initially, the homologous RNA sequences from diverse organisms are aligned based on similarities in their primary sequence.[15] These conserved sequence sets are used to align the more variable regions of the sequences. Subsequently, the base sequences are searched for covariation as possible pairing partners.[15] Developing secondary structure models from these alignments requires minimizing the free energy associated with pairing interactions.[15] Moreover, it is assumed that the free energy of each base pair is independent of all other pairs within the same predicted structure. This assumption, more commonly referred to as the Tinoco-Uhlenbeck postulate, states that the total free energy is the sum of all of the base pair free energies.[24] Dynamic programming then analyzes the ways that the base pairs can be constructed on an RNA strand and constructs a dot plot that produces a graphical energy plot for a given sequence.[24] This dot plot represents the lowest free energy for a structure that contains the pairing and provides a picture of all alternative structures.[24] Thereafter, a structure with the lowest total free energy is selected as the final prediction of RNA secondary structure.[24] However, this method is largely manual and requires significant user input, requires numerous homologous sequences that can be well aligned, and is limited in its ability to find non-canonical base pairings due to restrictions in

the dot plot construction.[31] However, since it predicts an approximate 98% of secondary structure base pairings and some tertiary pairings in crystal structures for well-aligned RNAs, it is referred to as the "gold standard" for RNA structure prediction.[31]

Free Energy Minimization Methods

Secondary structures in this method are computed by minimizing the total free energy of the individual substructures, including stems, loops, and bulges.[1] This thermodynamic approach can be applied to a single RNA sequence or functionally similar RNA sequences.[1] Out of all complementary sequence choices, the most energetically stable molecules are chosen. The folding of a primary sequence into loops include bases that are bonded which stabilize the RNA and have negative free energy, as well as unpaired bases that form destabilizing loops and have positive free energy.[1,11] Moreover, hairpin, interior and bulge loops destabilize energies.[1,11] Stems can include base pairing interactions or base stacking interactions, the dipole-dipole and van der Waals forces between bases.[1,11] Base stacking, as opposed to base pairing, is characterized with higher free energy.[1,11] Free energy tables, that denote the relative energy associated with specific base pairing interactions, are used to calculate the total free energy of an entire structure.[22] In addition, similar to the comparative analysis method, energy dot plots are constructed.[22] This method assumes that the most likely structure is identical to the energetically preferable structure.[22] In contrast to comparative analysis, this approach does not require prior sequence alignment. Also, multiple software programs, including Mfold and RNAfold, increase automation with relation to the comparative analysis technique.[11,33] However, this method is not without limitations. Tertiary interactions can affect the total free energy but are not incorporated into the free energy tables and are therefore ignored in calculations.[16] Moreover, the correct substructure may not be the structure associated with optimal free energy.[16] Therefore, multiple

suboptimal folds must also be considered and analyzed as possible candidates.[16] This approach is capable of predicting secondary structures with an accuracy as high as 73%.[16]

SHAPE and SHAPE-MaP

Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) has emerged as a more robust way of mapping RNA secondary structures with single-nucleotide resolution.[7] In this method, the RNA of interest is modified with a SHAPE reagent, (N-methylisatoic anhydride (NMIA) or a related molecule with similar properties, an electrophile that acylates conformationally flexible 2'-hydroxyl groups.[7,25] Local flexibility, as an analytical tool, provides information about the sequence, structure, and biological function of an RNA.[7,25] When reverse transcribed, these additions cause early termination resulting in multiple cDNA fragments with lengths corresponding to the location of the flexible 2'-hydroxyl groups.[17] Subsequently, electrophoresis using fluorescent-labelled primers separates the various fragments according to fragment size.[17] The complete SHAPE method includes an experimental set, a control set and at least one sequencing ladder.[17] The fragments from all sets, following electrophoresis, are analyzed to calculate the SHAPE reactivity of each nucleotide using the software program ShapeFinder.[17] The SHAPE reactivity refers to the relative stability of the nucleotide and can be converted to ΔG_{SHAPE} energy terms which are used in the RNAstructure program to provide accurate, secondary structure models for the RNA.[17] Similar to the SHAPE technique, a recent high throughput method known as SHAPE-MaP enables analysis of low-abundance RNAs and structure prediction of transcriptome-wide systems.[32] Selective 2' hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) yields high-resolution models, and predicts elements such as pseudoknots that could not have previously been analyzed with the SHAPE technology. Similar to SHAPE, SHAPE-MaP utilizes purified, folded RNA and an

electrophile such as 1-methyl-7-nitroiasatoic anhydride (1M7), a derivative of NMIA, or any other equivalent molecule.[32] 1M7 probes for and forms adducts with conformationally flexible ribose 2'-hydroxyl groups in the RNA.[32] A complete SHAPE-MaP experiment requires analyzing three distinct tests, including two control reactions and an experimental reaction.[32] The two control reactions include a DMSO control and denaturing control (DC). In the DMSO control, the SHAPE-MaP reagent 1M7 is not added to the folded RNA rather it is dissolved in DMSO, a polar aprotic solvent, only.[32] This control reaction will measure the intrinsic background mutation rate of the reverse transcription reaction and detect naturally occurring RNA modification events.[32] In the DC reaction, the RNA is suspended in a denaturing buffer that contains formamide and is incubated at 95 °C before modification with the SHAPE reagent.[32] During this control reaction the nucleotides are modified relatively evenly which permits the analysis of sequence and structure specific biases in detecting the adduct-induced mutations.[32] In the experimental reaction, the folded RNA will be incubated in modification buffer and the 1M7 reagent will be subsequently added.[32] These 1M7 induced adducts are incorporated into the RNA of interest which is subsequently reverse transcribed in the presence of Mn^{2+}, causing misincorporation at adduct sites and induction of mutations in the cDNA transcripts.[32] This mutational profiling aspect is relatively efficient with 50% of the induced adducts detected as mutations in the cDNA transcripts.[32] Large RNAs, such as those of betacoronaviruses, benefit from random priming in order to facilitate even coverage of the genome.[32] The mutation-prone cDNA transcripts are then used to construct a cDNA library for Illumina sequencing.[32] Deep sequencing of this cDNA library provides us with an avenue to obtain biochemical data to create RNA secondary structure models using SHAPE reactivity, Shannon entropy, and pairing probability models. A software pipeline of programs including ShapeMapper and RNAStructure

are used in the prediction of secondary structure models.[32] Further details regarding the precise methodology of SHAPE-MaP, such as biochemical probing and analysis, will take place in the subsequent methods section.

Therefore, it can be hypothesized that RNA secondary structures, visualized using the SHAPE-MaP methodology, conserved amongst MHV-A59, BCoV, and MERS-CoV are likely candidates for viral replication, specifically those flanking transcription regulatory sequence (TRS) regions.

CHAPTER II

METHODS

Growth of Cells and Virus

The MHV-A59 strain used, MHV-A59 1000, is a recombinant virus developed via a reverse genetics approach originally described by Yount et. al.[36] Additionally, BCoV (strain Mebus) viruses were received from the American Type Culture Collection (ATCC) and MERS-CoV was grown in the laboratory of Dr. Chien-Te K. Tseng at the University of Texas Medical Branch in Galveston, TX. The viruses under study were cultured and maintained in various cell lines. MHV-A59 was grown in DBT cells, BCoV was cultured in Human Rectal Tumor (HRT) cells, and MERS-CoV was maintained in Vero E6 cells. Moreover, the various cell lines were passaged in T175 flasks. Nicely confluent cells, estimated at 3×10^7 cells/flask, were infected with their respective virus stock. The cells were washed with 1 X Dulbecco's Modified Eagle's Medium (DME 0) and infected at a multiplicity of infection (MOI) of 0.1, or approximately 200 µL of virus stock, diluted in 2.5 mL of DME 2 per flask. Subsequently, 2.5 mL of the virus infused DME 2 solution was added to each flask and rocked for 60 minutes. Afterwards, 20 mL of DME 2 is added into each flask and incubated at 37 ˚C for 2 days. When 80% of the virus-infected cells detach from the surface of the flask and the remaining 20% of the infected cells are in syncytia, the virus is clarified. This clarification process requires centrifugation in a Jouan CR 412 Benchtop Refrigerated Centrifuge at 4000 RPM for 90 minutes.

Viral Quantification via Plaque-Based Assays

In order to determine the concentration of virus in the cell samples, a plaque assay is conducted. A 6-well plate containing the respective cell line is cultured at 1×10^6 cells/well.

After the cells are confluent, they are washed with 1 X DME 0. A sample of clarified virus is diluted to 10^{-4}, 10^{-5}, and 10^{-6} and 200 µL of each dilution is added in replicate to the 6-well plate and rocked for 60 minutes. The cells are then overlaid with 2.5 mL of agarose and 2 X DME2 mix at 37˚C for 2 days. After removing the agarose, the cells are stained with crystal violet and plaque-forming units are manually counted to determine the viral titer.

Viral RNA Purification and Extraction

Infected cell culture fluids are pooled after defrosting and clarified in the Jouan CR 412 Benchtop Refrigerated Centrifuge at 5000 g for 60 minutes. Subsequently, 36 mL of virus supernatant is pipetted into each of the six Ultra-Clear SW 28 ultracentrifuge tubes and underlaid with 1.5 mL of 30% (w/w) sucrose in MSE using a Pasteur pipet. It should be noted that the tubes in their buckets were weighed and are within 100 mg of each other. Subsequently, suspensions were pelleted for 2.5 hours at 25,000 RPM (112,5000 x g) in a SW28 rotor using a Thermo Sorvall WX 100+ Ultracentrifuge. The supernatant was poured off and the tubes blotted dry, with no residual liquid in the tube so as to keep the volume as small as possible. Another 36 mL of virus supernatant were pipetted into each SW 28 tube and the process was repeated twice for a total of three spins with the Thermo Sorvall WX 100+ Ultracentrifuge. Afterwards, the pellet in one of the 6 SW 28 tubes was resuspended in 200 µL of MSE buffer and transferred from tube to tube, ensuring that all pellets have detached from the surface of their tube. In order to maximize consistency amongst the samples, pellets were dispersed via sonication in the Heat Systems Ultrasonics Inc. Sonicator XL-2020. The virus suspensions, located in sealed tubes, were dispersed in an ice water bath using three bursts of sonication lasting 20 seconds, each at 100 W with 20 second breaks in between. The pooled, sonicated virus is then overlaid on 11.6 mL 20-60% (w/w) sucrose in MSE gradient and centrifuged in an SW 41 rotor in the Thermo

Sorvall WX 100+ Ultracentrifuge at 25,000 RPM overnight. The visible opalescent virus band in the middle of the tube is then collected by puncturing the side of the tube with a 20-guage needle attached to a 3-mL syringe. After collecting the band, the refractive index of the virus was determined using a Bausch & Lomb Abbe-3L Refractometer to ensure the identity of the purified sample. The refractive index of the sample was used to measure a buoyant density that could be equated to the buoyant density of betacoronaviruses, which is approximately 1.17-1.19 g/cm^3. After confirming the identity of the purified virus, the sample was diluted with enough MSE buffer to reach a volume of 11.5 mL to fill an Ultra-Clear SW 41 ultracentrifuge tube. 11.6 mL of MSE buffer was added to another two balance tubes and weighed. The diluted virus was then spun in a SW 41 rotor using the Thermo Sorvall WX 100+ Ultracentrifuge at 35,000 RPM for 1 hour at 4˚C. The pellet was resuspended in 100 µL of viral lysis buffer and the virions are lysed by the addition of 17.2 µL of 10% (w/v) SDS and 1.5 µL of 100 µg/mL proteinase K. The virus was incubated at 25 ˚C, or room temperature, for 30 minutes and viral RNA was extracted three times with phenol: chloroform and once with only chloroform. In order to concentrate the viral RNA, the sample was ethanol precipitated overnight.

Subjection of Viral RNA to SHAPE-MaP conditions

After recovering the ethanol precipitated sample, the viral RNA was resuspended in 10.1 µL of warm modification buffer. In order to determine the concentration of viral RNA extracted, the sample was quantitated on a Thermo Fisher Scientific NanoDrop spectrophotometer using the RNA setting. In order to conserve the sample, a 1:10 dilution of the 1µL RNA was conducted using warm modification buffer. Subsequently, the RNA was aliquoted into three equal reactions: a 1M7 reaction, a DMSO control reaction, and a denatured control reaction. The DMSO control is used to measure the intrinsic background mutation rate of the reverse

transcription reaction, described later. The denatured control reaction ensures that the nucleotides are modified evenly while taking into account any site-specific or sequence-specific biases in detecting mutation rates. These separate reactions were run in parallel with varying experimental conditions. A minimum of 3.37 µg per reaction was used; however, if more RNA was present it was equally distributed amongst the three sets. After the samples were dried in the Thermo Savant SC110A Speed Vac, the Bio-Rad PTC-100® Thermal Cycler was pre-warmed to 95 ˚C. The aliquoted RNA for the 1M7 experimental reaction and DMSO control reaction was resuspended in 10 µL modification buffer and incubated at 37 ˚C for 15 minutes. In addition, the aliquoted RNA in the denatured control reaction was resuspended in 10 µL of denaturing buffer, containing formamide, and incubated at 95 ˚C for 2 minutes in the Bio-Rad PTC-100® Thermal Cycler. The following samples were also pre-warmed at 37 ˚C: 100 mM 1M7 in DMSO, 3 µL of 50 mM EDTA, and 100 µL of DMSO. To both plus-reagent reactions (1M7 and denatured control), 1.1 µL of the pre-warmed 100 mM 1M7 in DMSO was added. To the DMSO control reaction, 1.1 µL of pre-warmed DMSO was added. All samples were incubated at 37 ˚C for 70 seconds and 1.1 µL of 50 mM EDTA was added to stop all three reactions. Afterwards, all samples were ethanol precipitated overnight after the addition of 1 µL of glycogen as a carrier.

Reverse Transcription of Modified RNA

After recovering the ethanol precipitated samples, each pellet was dissolved in 18.4 µL of RT buffer, which includes 0.7 mM dNTPs, 50 mM Tris HCl, 75 mM KCl, 6 mM $MnCl_2$, and 14 mM DTT, for 15 minutes. Subsequently, 0.70 µL of Random Primer 9 (300 ng/ µL) from New England Biolabs, a 9-mer that ensures even coverage of the viral RNA, and 1 µL of SuperScript II RT (200 U) was added and the mixture was incubated at 42 ˚C for 3 hours. The samples were subsequently ethanol precipitated overnight. As a result of this process, the positions and

frequencies of the SHAPE adducts are represented by mutations in the cDNA transcripts. This profiling is capable of detecting approximately 50% of SHAPE adducts as mutations.

Second Strand Synthesis of Modified cDNA Transcripts

Subsequently, the Bio-Rad PTC-100® Thermal Cycler was pre-heated to 70 ˚C and the ethanol precipitated samples were recovered. After ensuring that the samples were completely dry, the pellets were resuspended in 20 µL of RNAse/DNAse free H_2O. A 1:10 dilution of the samples, in DEPC H_2O, was then quantitated on a Thermo Fisher Scientific NanoDrop spectrophotometer using the DNA setting. The appropriate volume containing 100 ng of cDNA from each experimental reaction, the amount required for second strand synthesis, was placed in the pre-heated thermal cycler at 70 ˚C for 15 minutes. After precooling the Bio-Rad PTC-100® Thermal Cycler to 16 ˚C, second-strand synthesis with the NEBNext® Ultra II Non-Directional RNA Second Strand Synthesis Module was performed in 48 µL of DEPC H_2O, 8 µL of 10X Second Strand Synthesis Reaction Buffer and 4 µL of Second Strand Synthesis Enzyme Mix added to the heat inactivated first-strand synthesis reaction. This mixture was then incubated at 16 ˚C for 2.5 hours in the pre-cooled thermal cycler. These double-stranded cDNA transcripts were subsequently sent to the AgriLife Genomics and Bioinformatics Service and Texas A&M Institute for Genome Sciences and Society for Illumina sequencing. A Nextera XT DNA Library Preparation Kit was used to prepare the three cDNA libraries. Subsequently, quality control measurements were conducted in order to ensure that correctly sized cDNA transcripts were generated. These transcripts were then sequenced on the Illumina MiSeq platform, generating FASTQ-formatted output files. Further details on how the bioinformatics data was analyzed are described next. It should be noted that this procedure was repeated twice for each of the viruses analyzed.

SHAPE-MaP Analysis

In order to determine depth of sequencing, a series of Unix shell scripts were written. 6 zipped FASTQ files were divided into 3 groups of 2 paired end reads, each group containing information of DMSO, DC, and 1M7 sequence data. Each group of 2 files is then inputted into manalign.sh that utilizes Bowtie 2 to align the cDNA sequences relative to their respective reference sequences, obtained from the NCBI databases. This script provides .sam output files that are then converted into .bam files using the samtobam.sh script. The compressed .bam output files are binary counterparts of the .sam text files and are suitable for sorting. These .bam files are then sorted along the genome using the sortbam.sh script. The output file from this script is then inputted into the depthfrombam.sh script which provides the depth of sequencing at each nucleotide position that can be visualized in Microsoft Excel. An approximate 5,000 reads across the genome are recommended in order to gather high-resolution structural information. Figure 3 provides a graphical representation of this flow.

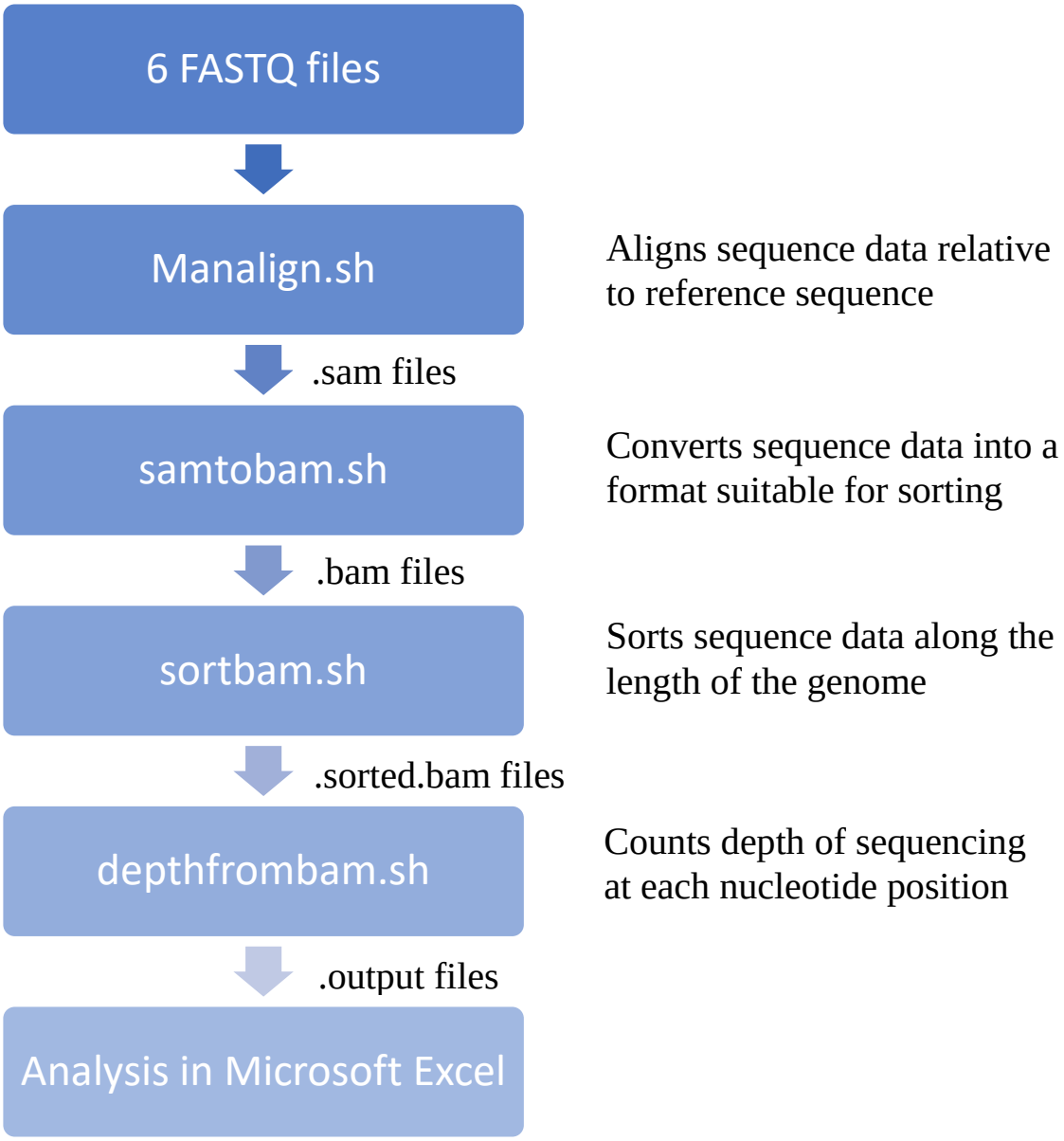

Figure 3. Flowchart indicating the scripts and descriptions involved in analysis of depth of sequencing

After ensuring that adequate depth of sequencing was achieved, the SHAPE-MaP pipeline was run to calculate a .ct or connectivity file that provides base pairing information, described more thoroughly shortly. The original 6 FASTQ files were unzipped and the file names were entered into configuration files which specify which virus the RNA sequences belong to and separates the three experimental reactions (1M7, DMSO, and DC) for each virus.

These configuration files, specifying the input file, output file, and reference sequence for each of the three groups, are called by the ShapeMapper.py script. This script aligns and trims the sequence reads relative to the reference sequence for each respective virus using the same Bowtie 2 script previously described to determine depth of sequencing. The mutation frequency of each nucleotide position is also calculated by the ShapeMapper.py script. Frequency is calculated by subtracting the intrinsic mutation rate of the DMSO control reaction from that of the 1M7 experimental reaction. The pseudo-free energy term is then calculated by ShapeMapper.py using the mutation frequency, generating a .shape output file, which contains the nucleotide position and the SHAPE reactivity encoded at that region, partition function modules, and Shannon entropy values. Thus, areas of low SHAPE reactivity signify regions of relatively low mutation rates and low pseudo-free energies. Shannon entropy is calculated for each nucleotide using the partition function module. It provides a measure of possible alternative structures that can be formed given the same folding parameters. Therefore, areas of low Shannon entropy indicate that a highly structured, predominant secondary structure is likely present. The pseudo-free energy values are then passed onto RNAStructure, developed by the Mathews Lab, which generates the .ct files, containing information regarding nucleotide position, nucleotide base sequence, and nucleotide base pairing interactions. This software slides along the genome using a series of 3,000 nucleotide sliding windows each of which is offset from the previous window by 500 nts to generate a series of overlapping 3000 nt individual .ct output files for each window. The .shape output file and individual .ct output files are then called upon by SuperFold.py which combines the input .ct files into a larger merged .ct file, displaying the degrees of connectivity amongst the entire genome while also generating arc diagrams representative of base pairing probabilities. An 80% probability of base pairing is indicative of

highly probable interaction, as indicated by the legend in Figures 5 and 6. The arc diagrams of secondary structure of the two viral RNA genomes was inspected for regions of low Shannon entropy. Potential conserved structures in these regions were identified and outlined by red boxes in Figures 5 and 6. These possible conserved structures were then visualized at the nucleotide level as explained in detail below. Figure 4 provides a graphical representation of this flow.

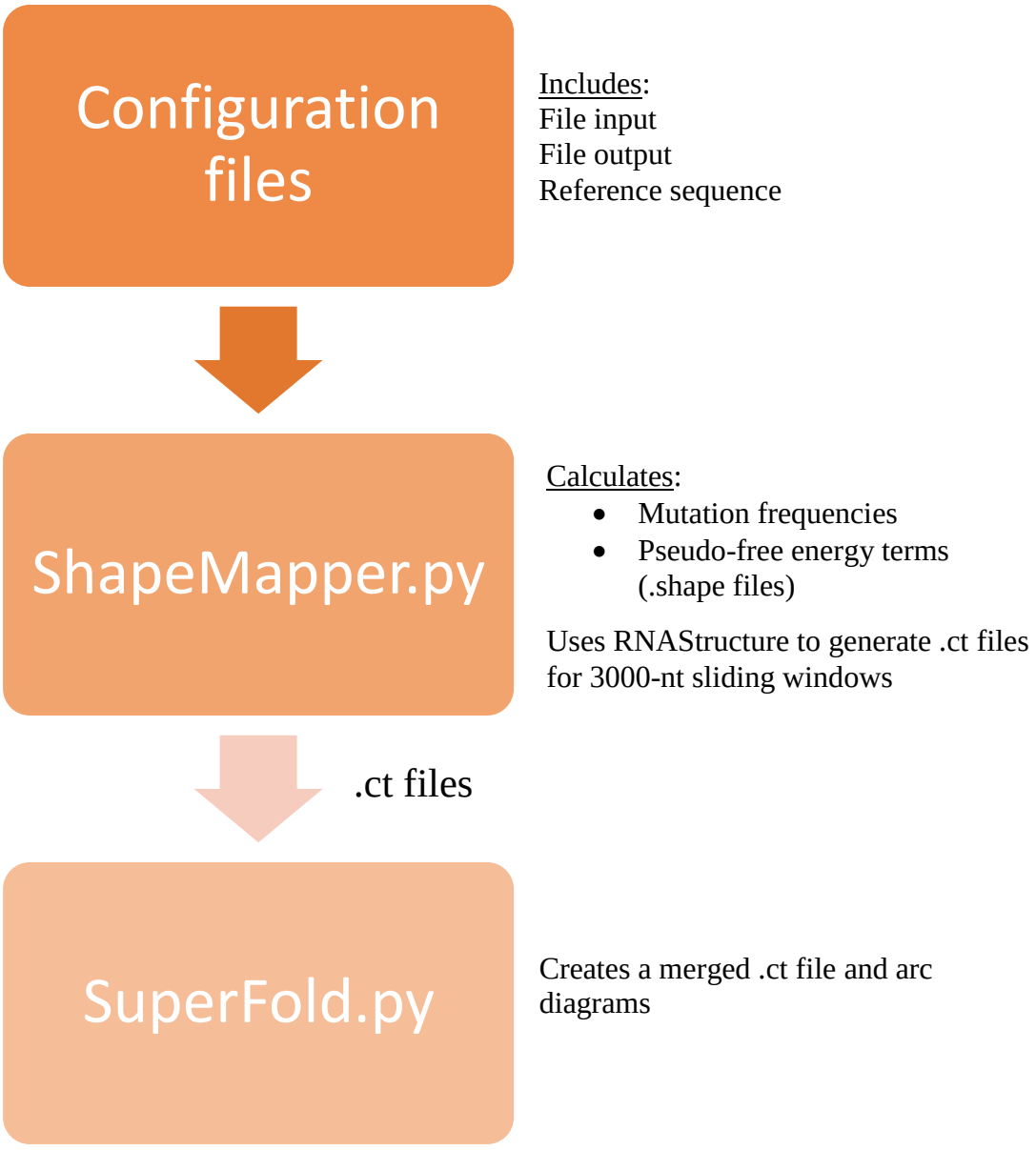

Figure 4. Flowchart indicating the scripts and descriptions involved in secondary structure visualization

With the assistance of Dr. Byung-Jun Yoon, the merged .ct files for candidate conserved structures were extracted using test_read_ct.R and stored into separate .ct files using test_react_ct_batch.R, described in detail in the appendix section. These extracted files are converted into dot bracket (.db) files, another form of displaying nucleotide sequences and their interactions using dots and parenthesis to discriminate between base-pairing and non-base-pairing nucleotides, using the ct2dot server developed by the Mathews Lab. These dot bracket files were then visualized using FORNA, an RNA secondary structure visualization platform, provided by the ViennaRNA Web Service. The Results section will go into further detail on which regions were specifically visualized and why.

CHAPTER III

RESULTS

Depth of sequencing and full genome analysis via SHAPE-MaP

The average depth of sequencing achieved for MHV and BCoV was 5350 reads, meeting the 5000 average reads requirement for SHAPE-MaP implementation.

As described under Methods, SHAPE reactivity profiles, Shannon entropy profiles, and arc diagrams depicting base pairing probabilities are generated by the SHAPE-MaP pipeline, specifically ShapeMapper and SuperFold. Figures 5 and 6 include profiles for MHV-A59 and BCoV. Note that in the figures, the red-colored graph represents Shannon entropy and the arc diagrams are situated below the nucleotide position:

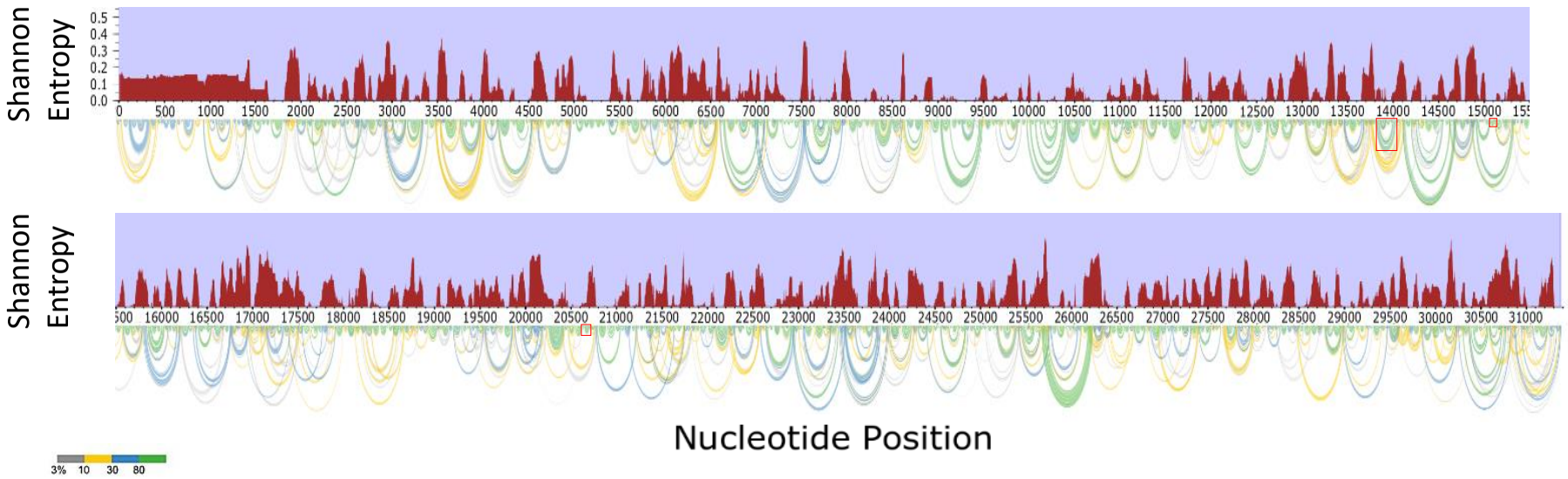

Figure 5. Genome-wide analysis of MHV, including Shannon entropy and base pairing

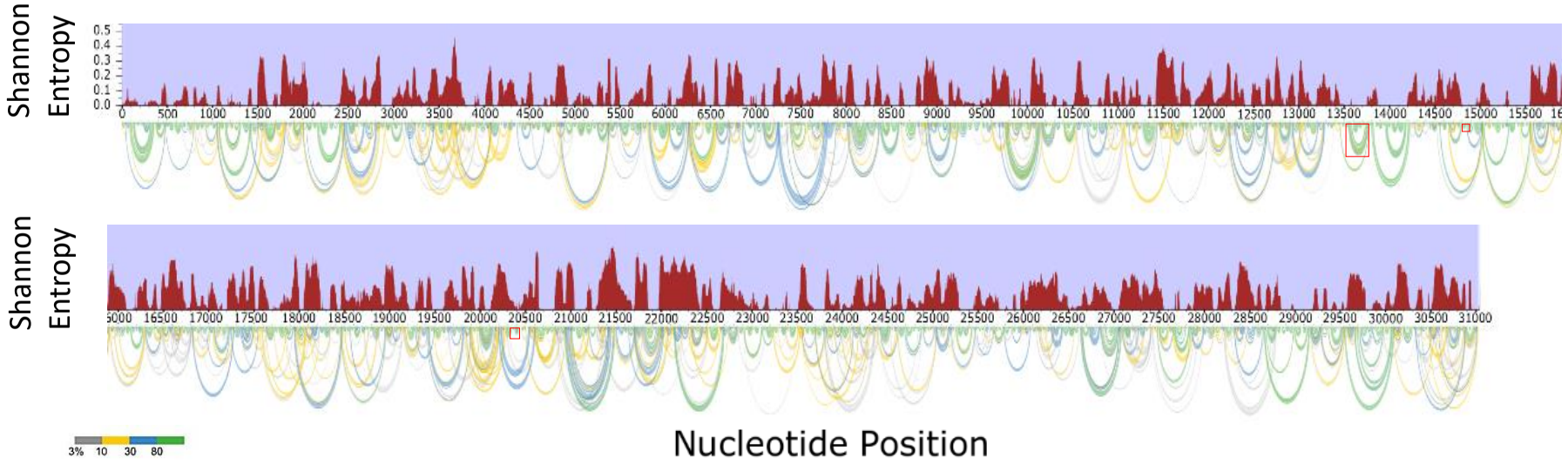

Figure 6. Genome-wide analysis of BCoV, including Shannon entropy and base pairing

Isolating Areas of Interest

Prior to scanning for regions of interest, sequence alignment was conducted using the LALIGN program that is part of the EMBOSS suite of programs. This software searches for similarities in sequences between MHV-A59 and BCoV and permits visualization of gaps where one or more nucleotides have been deleted in the sequence. Such an alignment was conducted separately for orf 1a, orf 1b, and structural and accessory protein coding regions 3' of orf 1b (about 10,000 nts for each virus) and the 3'-adjacent 3'UTR.

Subsequently, manual analysis to identify regions fulfilling the following criteria was conducted:

a. Areas of low SHAPE reactivity or more highly structured regions
b. Areas of low Shannon entropy likely to contain more robust predictions of structured regions,
c. Regions containing an 80% or greater probability of base pairing

Table 1 lists the sequences identified:

Table 1. Areas of interest isolated based on comparison of arc diagrams and relative levels of Shannon entropy

Nucleotide Start Position	Nucleotide End Position	Virus	Shannon Entropy	Comparison of Arcs
		MHV	-1	
2200	2325	BCoV	-2	Similar
		MHV	1	
2700	2750	BCoV	-2	Similar
		MHV	-2	
3600	3700	BCoV	2	Similar
4200	4550	MHV	-2	
4050	4400	BCoV	-2	Similar
4700	4800	MHV	-1	
4600	4700	BCoV	-1	Similar
5690	5750	MHV	-1	
5475	5600	BCoV	-1	Similar
6700	6900	MHV	2	
6450	6600	BCoV	1	Similar
8320	8500	MHV	1	
8050	8250	BCoV	-1	Similar
9050	9500	MHV	1	
8780	9250	BCoV	-2	Similar
10400	10550	MHV	-2	
10150	10500	BCoV	-1	Similar
11470	11675	MHV	-2	
11200	11400	BCoV	-1	Similar
12250	12650	MHV	1	
11900	12300	BCoV	-2	Similar
13700	14000	MHV	-1	
13550	13800	BCoV	-2	Similar
15050	15280	MHV	2	
14800	15050	BCoV	-1	Similar
15850	16020	MHV	1	
15600	15750	BCoV	-2	Similar
16400	16520	MHV	-1	
16100	16300	BCoV	-1	Similar
17900	18200	MHV	-2	
17600	17850	BCoV	1	Similar
19300	19500	MHV	-1	
18850	19050	BCoV	2	Similar
20250	20750	MHV	-1	
20000	20450	BCoV	-2	Similar
20800	21000	MHV	-1	
20500	20700	BCoV	-2	Similar
22100	22200	MHV	-2	
21800	21900	BCoV	-2	Similar
24530	24800	MHV	-1	
24400	24630	BCoV	-2	Different
25700	26300	MHV	1	
N/A	N/A	BCoV	2	Different
26050	26550	MHV	-2	
25550	26000	BCoV	-2	Different

Table 1. Areas of interest isolated based on comparison of arc diagrams and relative levels of Shannon entropy (continued)

26840	27080	MHV	1	
26700	27000	BCoV	1	Similar
26650	26760	MHV	1	
26500	26600	BCoV	-1	Similar
27950	28050	MHV	-2	
28000	28100	BCoV	2	Similar
29350	29550	MHV	1	
29080	29300	BCoV	2	Similar
29650	29800	MHV	1	
29400	29500	BCoV	2	Similar
31000	31300	MHV	1	
30650	30900	BCoV	1	Similar
30200	30800	MHV	-1	
29500	31000	BCoV	-1	Similar

It should be noted that a relative number (-1, 0, 1) was used to assess the level of Shannon entropy, with -1 representing low Shannon entropy and +1 representing high Shannon entropy. Arc diagrams were also visually assessed and evaluated as either "similar" or "different."

Visualization of Secondary Structures

With the assistance of Dr. Byung-Jun Yoon, the .db files for each region of interest, generated from the merged .ct file, was fed into the FORNA application, generating RNA secondary structure predictions. The BCoV and MHV-A59 structure predictions for each area of interest were manually compared using the alignment mentioned previously with LALIGN. From 31 areas of interest, 3 areas of interest containing conserved secondary structure predictions were identified. These structures are shown in Figures 7-11.

The models were visually inspected for sequence variations between the two viruses despite displaying similar configurations. The base pairs highlighted in pink display such base pairs. Particular note should be taken of covariation on both sides of a stem as this provides evidence for conservation of the structure over evolution. Moreover, the arc diagrams of

secondary structure corresponding to the isolated, conserved areas of interest are outlined in red boxes on the base pairing probability models in Figures 5 and 6.

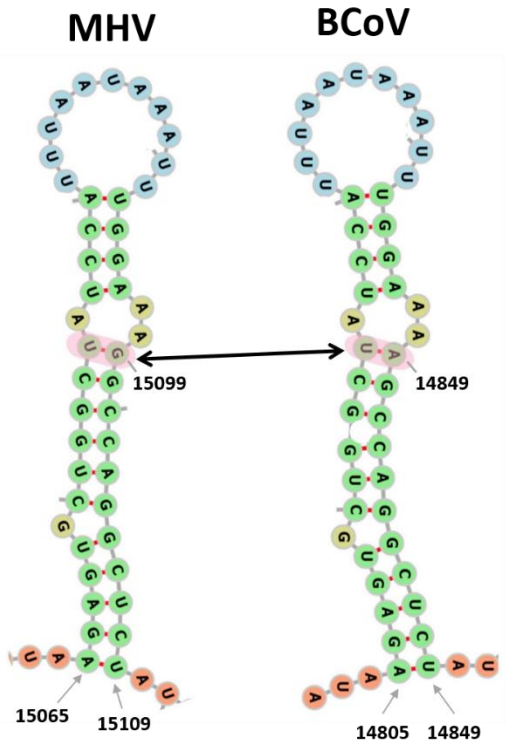

Figure 7. Comparison of conserved stem loop in MHV-A59 and BCoV

In the above figure, sequence comparison between MHV-A59 and BCoV reveals a single point mutation. As highlighted, guanine (G) at nucleotide position 15099 in MHV-A59 is converted to adenosine (A) at nucleotide position 14849 in BCoV. This mutation; however, does not change the overall configuration of the conserved stem loop.

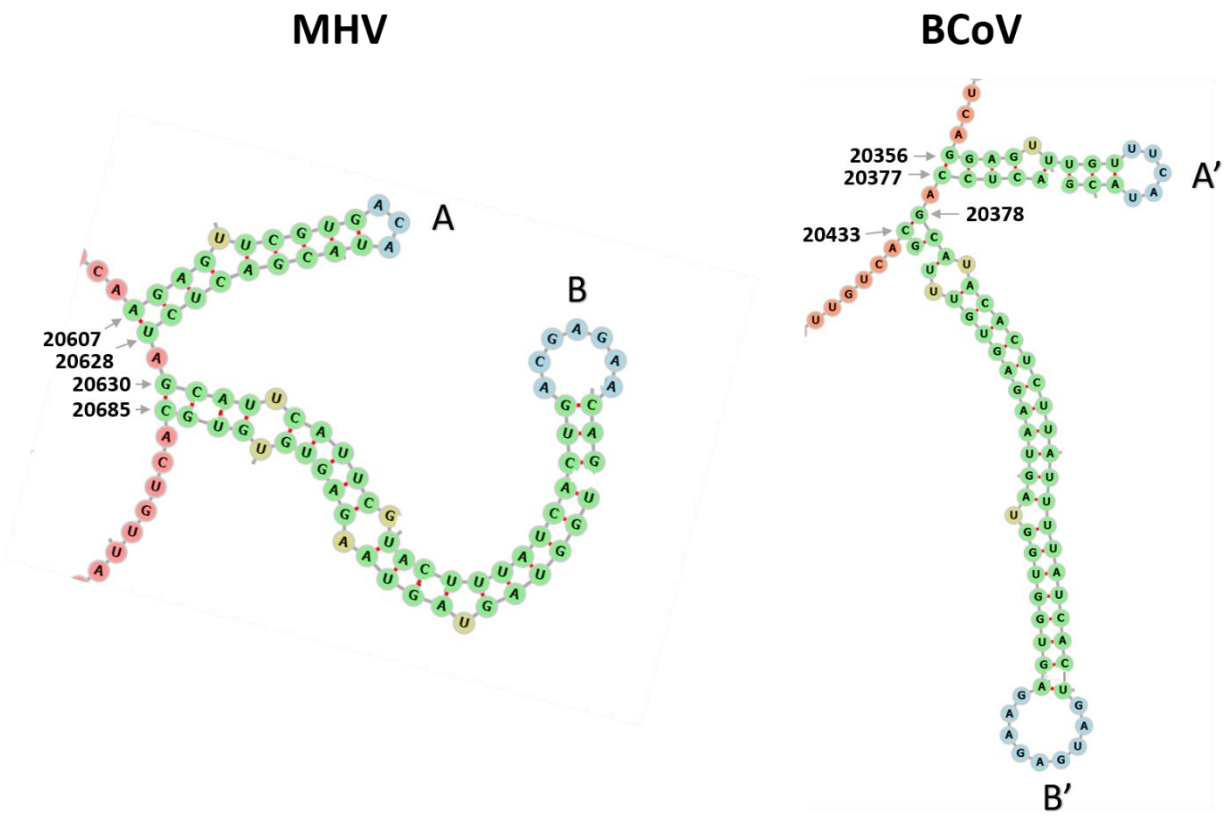

Figure 8. Comparison of two conserved stem loops in MHV-A59 and BCoV

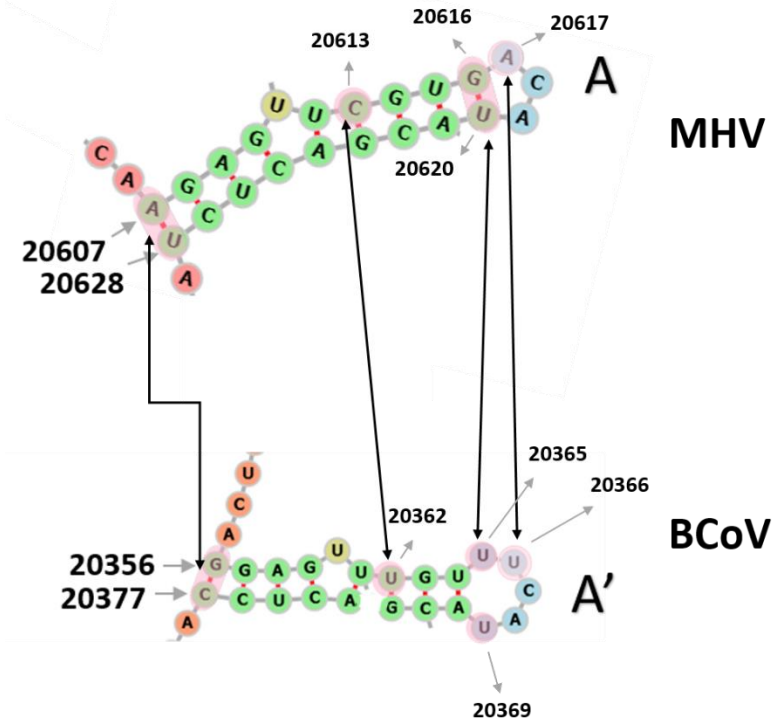

Figure 9. In depth comparison of stem loops A from MHV-A59 and B' from BCoV

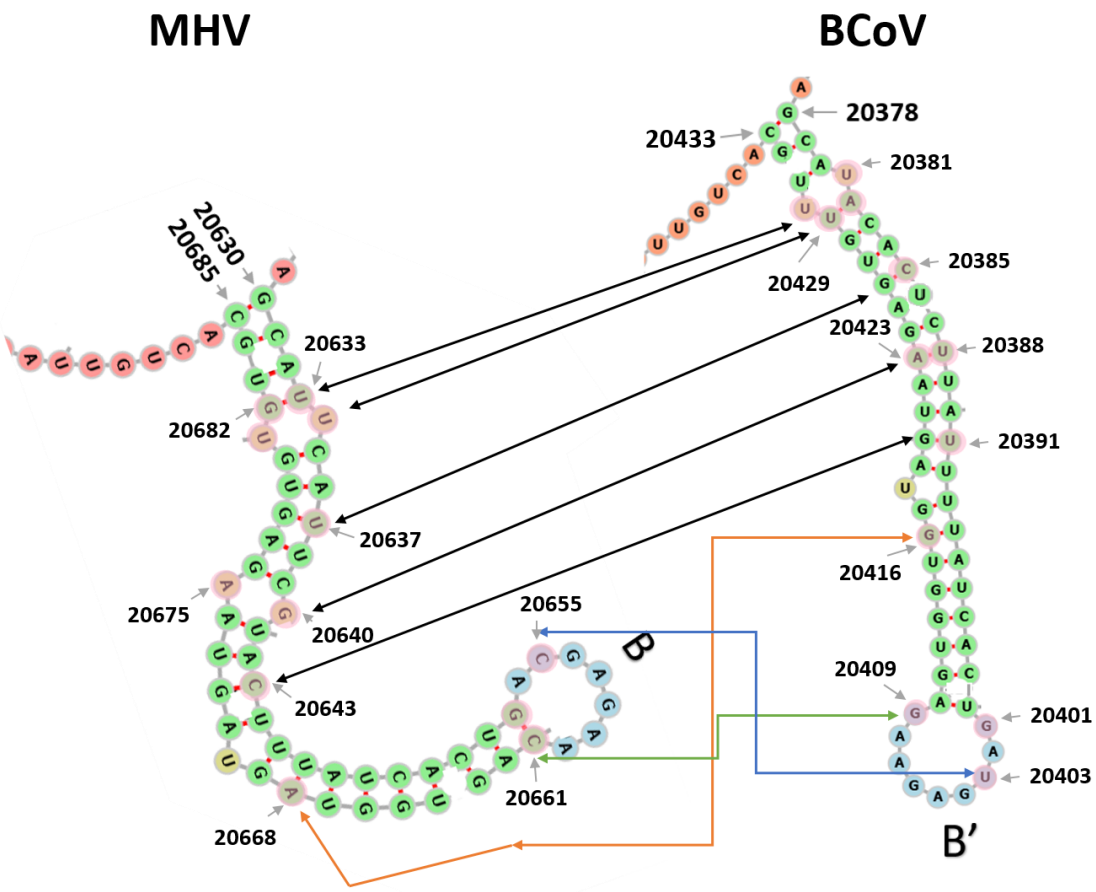

Figure 10. In depth comparison of stem loops B from MHV-A59 and B' from BCoV

Figure 8 depicts the conservation of two stem loops in both viruses, labeled A and B in MHV-A59 and A' and B' in BCoV. An A-U base pair at nucleotide positions 20607 and 20628 in MHV is converted to a G-C base pair at nucleotide positions 20356 and 20377 in BCoV. This results in the formation of a highly stable G-C pairing interaction at the base of the stem and could provide support for covariation. In addition, a point mutation at nucleotide position 20613 in MHV-A59 converts cytosine (C) to uracil (U) at nucleotide position 20362 in BCoV. Thus, a stable G-C base pair observed in MHV-A59 is lacking in BCoV at the corresponding position. The free energy of the substructures were calculated and compared to measure variations in

stability. The free energy (ΔG) of substructure A in MHV-A59 was calculated to be -8.0 kcal/mole whereas the ΔG of BCoV was calculated to be -4.9 kcal/mole.

Figure 10 depicts several point mutations between structure B in MHV and structure B' in BCoV. At nucleotide position 20682 in MHV-A59, guanine (G) is converted to uracil (U) at nucleotide position 20430 in BCoV. Interestingly, the G-U base pair conversion in BCoV still induces a similar UU bulge located in MHV-A59. Moreover, the uracil (U) at nucleotide position 20664 in MHV-A59, a member of the UU bulge, is converted to adenosine (A) at nucleotide position 20382 in BCoV. This conversion establishes a U-A base pair that closes the UU bulge in BCoV. At nucleotide position 20637 in MHV-A59, uracil (U) is converted to cytosine (C) at nucleotide position 20385 in BCoV. This addition of a highly stable G-C base pair could further stabilize the stem loop in BCoV. In addition, at nucleotide position 20643 in MHV-A59, cytosine (C) is converted to uracil (U) at nucleotide position 20391 in BCoV. In this case; however, the more stable G-C base pair in MHV-A59 is mutated to a less stable G-U base pair in BCoV. Free energy calculations were conducted to determine the overall stability of the stem loop. The free energy (ΔG) of the stem loop in MHV-A59 was determined to be -6.5 kcal/mol and the free energy associated with its BCoV counterpart was determined to be -5.5 kcal/mol. At nucleotide position 20661 in MHV, cytosine (C) which forms a stable G-C base pair is converted to guanine (G) at nucleotide position 20409. This conversion contributes to the terminal loop, which is accordingly larger in BCoV than in MHV-A59 at this region

.

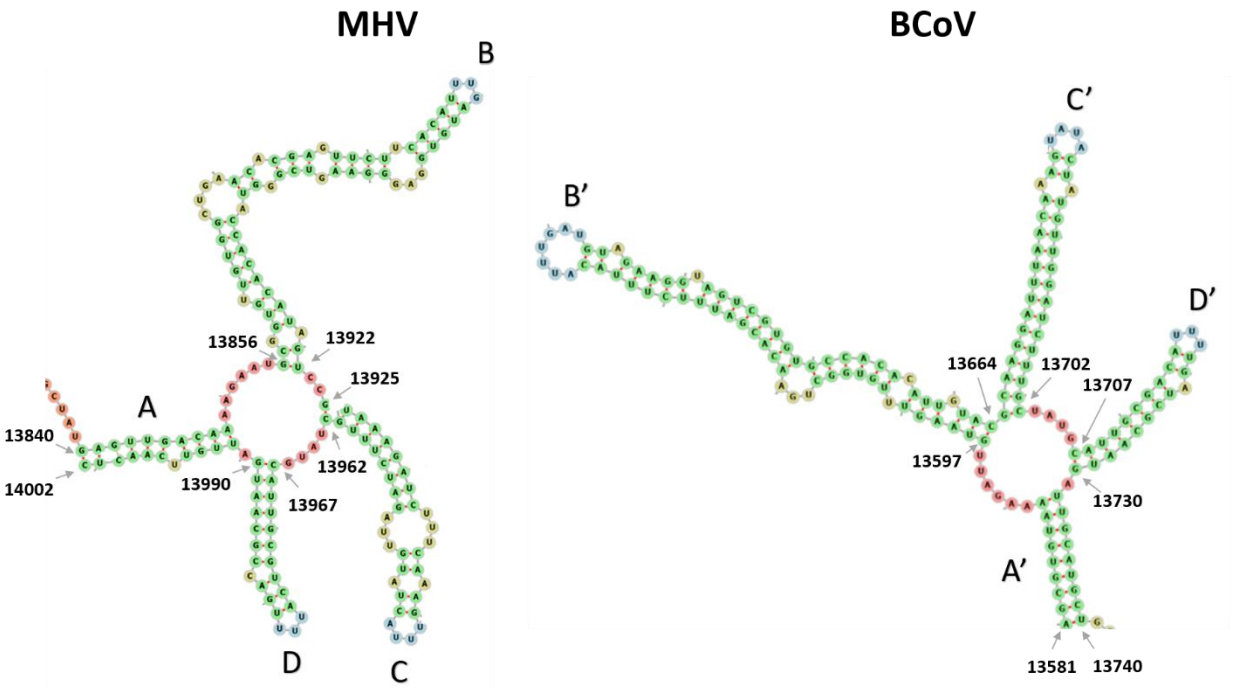

Figure 11. Comparison of multi-branched loop in MHV-A59 and BCoV

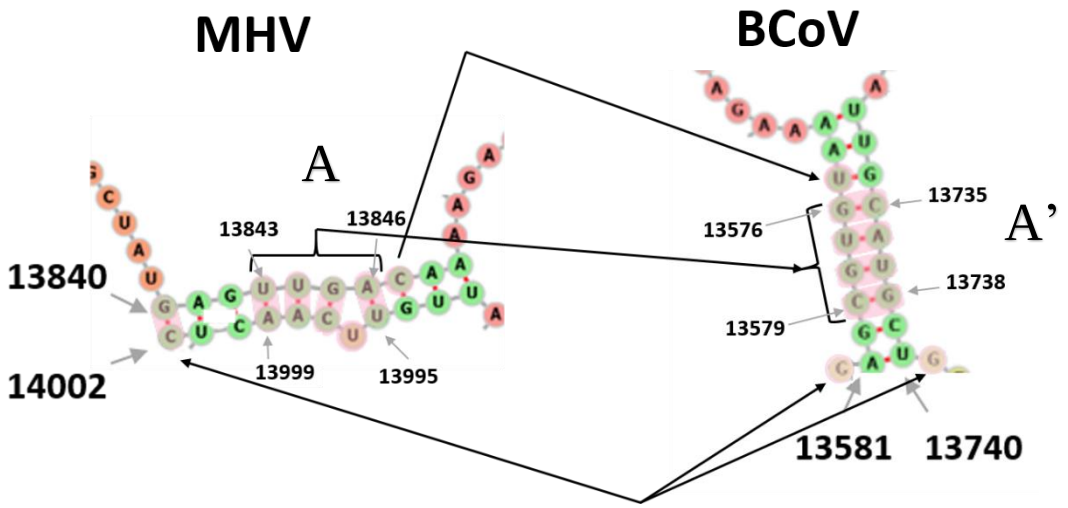

Figure 12. In depth comparison of stem loops A from MHV-A59 and A' from BCoV

The multi-branched loops depicted in Figure 11 have also been differentiated into separate substructures. The stem substructure A in MHV-A59 corresponds to the similarly folded stem substructure A' in BCoV. The three branches, B, C, and D, in MHV-A59 also similarly correspond to the three branches, B', C', and D', in BCoV.

While both stems (A and A') display similar conformations as depicted in Figure 12, divergences in nucleotide sequence should be noted. Cytosine (C) and guanine (G) at nucleotide positions 13580 and 13740, respectively, in BCoV combine to form the base of the stem in MHV, at nucleotide positions 13840 and 14002. This G-C base pair reduces the free energy associated with the stem and could increase the stability of the multi-branched loop. Interestingly, the central portion of the stem, outlined in brackets in Figure 10, is not conserved on a nucleotide-level. Rather there are complete base pair switches, including two U-A base pairs in MHV that are converted to G-C base pairs in BCoV, and the presence of a single-nucleotide bulge in MHV that is not present in BCoV. The base pair switches could also provide support for covariation. Thermodynamic calculations; however, yielded similar energy values for substructures A and A'. The free energy associated with substructure A in MHV-A59 is -4.2 kcal/mole, whereas substructure A' in BCoV is -3.9 kcal/mol. Similar analysis of substructures B, C, and D in MHV-A59 with relation to substructures B', C', and D' in BCoV is currently underway.

CHAPTER IV

CONCLUSION

Identification of Conserved RNA Secondary Structures

The very large size of coronavirus genomes presents a particular challenge for determination of RNA secondary structure. Initial attempts at using the SHAPE-MaP technique to visualize RNA secondary structure models for betacoronaviruses did not yield sufficient depth of sequencing for accurate modeling. However, modifications in the purification protocol, which reduced extraneous DNA contamination, yielded higher, recommended read depths. This allowed for the construction of accurate, nucleotide-resolution RNA secondary structure predictions. To the best of our knowledge, these are the largest genomes to have been successfully analyzed using the SHAPE-MaP technique. This methodology allowed for the identification of several conserved secondary structures depicted in Figures 7-10 and described in the Results section.

Biological Significance of Conserved Structures

Conservation of the aforementioned structures despite variations in sequence amongst MHV-A59 and BCoV suggest functional roles for these structures in viral replication. As described previously in the Results section and displayed in Figures 5 and 6, red boxes were used to outline the arc diagrams corresponding to the genomic regions containing the aforementioned conserved secondary structures. All three of the secondary structure models are located in open reading frame (ORF) 1b. Thus, they might play a role in modulating translation and could serve as binding sites for host or viral proteins. However, further experimentation via site-directed mutagenesis is needed to determine their exact functional role in replication.

Limitations and Future Directions

While previous literature suggests the presence of two pseudoknots, one near the frameshift region, starting at approximately nucleotide 13700, and the other located at the distal 3’ end, we were unable to uncover such structures using the standard SHAPE-MaP protocol. Future directions will include implementing the ShapeKnots program, also developed by the Mathews Lab, to identify potential pseudoknots given sequence constraints. Moreover, the identified conserved structures will be subjected to site-directed mutagenesis for functional analysis.

REFERENCES

1. Andronescu MS, Pop C, Condon AE. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *Rna*. 2009;16(1):26-42. doi:10.1261/rna.1689910.

2. Beniac DR, Booth TF. Structural Molecular Insights into SARS Coronavirus Cellular Attachment, Entry and Morphogenesis. *Molecular Biology of the SARS-Coronavirus*. 2009:31-43. doi:10.1007/978-3-642-03683-5_3.

3. Brian DA, Baric RS. Coronavirus Genome Structure and Replication. *Current Topics in Microbiology and Immunology Coronavirus Replication and Reverse Genetics*.:1-30. doi:10.1007/3-540-26765-4_1.

4. Coronavirus Infections. Equine Infectious Diseases. 2007:184-185. doi:10.1016/b978-1-4160-2406-4.50023-5.

5. Coronavirus Replication and Reverse Genetics. *Current Topics in Microbiology and Immunology*. 2005. doi:10.1007/b138038.

6. Coronaviruses. Advances in Virus Research. 2016. doi:10.1016/s0065-3527(16)x0004-8.

7. Ding Y. Statistical and Bayesian approaches to RNA secondary structure prediction. Rna. 2006;12(3):323-331. doi:10.1261/rna.2274106.

8. Dinman JD. Programmed –1 Ribosomal Frameshifting in SARS Coronavirus. *Molecular Biology of the SARS-Coronavirus*. 2009:63-72. doi:10.1007/978-3-642-03683-5_5.

9. Eickmann M. Phylogeny of the SARS Coronavirus. Science. 2003;302(5650). doi:10.1126/science.302.5650.1504b.

10. Gao L. Molecular phylogeny of coronaviruses including human SARS-CoV. Chinese Science Bulletin. 2003;48(12):1170. doi:10.1360/03wc0254.

11. Geis M, Middendorf M. A Particle Swarm Optimizer for Finding Minimum Free Energy RNA Secondary Structures. *2007 IEEE Swarm Intelligence Symposium*. 2007. doi:10.1109/sis.2007.368019.

12. Goebel SJ, Taylor J, Masters PS. The 3 cis-Acting Genomic Replication Element of the Severe Acute Respiratory Syndrome Coronavirus Can Function in the Murine Coronavirus Genome. *Journal of Virology*. 2004;78(14):7846-7851. doi:10.1128/jvi.78.14.7846-7851.2004.

13. Guan B-J, Wu H-Y, Brian DA. An Optimal cis-Replication Stem-Loop IV in the 5 Untranslated Region of the Mouse Coronavirus Genome Extends 16 Nucleotides into Open Reading Frame 1. *Journal of Virology*. 2011;85(11):5593-5605. doi:10.1128/jvi.00263-11.

14. Imbert I, Ulferts R, Ziebuhr J, Canard B. SARS Coronavirus Replicative Enzymes: Structures and Mechanisms. *Molecular Biology of the SARS-Coronavirus*. 2009:99-114. doi:10.1007/978-3-642-03683-5_7.

15. James BD, Olsen GJ, Pace NR. [18] Phylogenetic comparative analysis of RNA secondary structure. *Methods in Enzymology RNA Processing Part A: General Methods*. 1989:227-239. doi:10.1016/0076-6879(89)80104-1.

16. Juan V, Wilson C. RNA Secondary Structure Prediction Based on Free Energy and Phylogenetic Analysis. Journal of Molecular Biology. 1999;289(4):935-947. doi:10.1006/jmbi.1999.2801.

17. Kennedy SD. NMR Methods for Characterization of RNA Secondary Structure. RNA Structure Determination Methods in Molecular Biology. 2016:253-264. doi:10.1007/978-1-4939-6433-8_16.

18. Liu D. Bovine Coronavirus. *Molecular Detection of Animal Viral Pathogens*. May 2016:317-322. doi:10.1201/b19719-37.

19. Liu P, Leibowitz J. RNA Higher-Order Structures Within the Coronavirus 5′ and 3′ Untranslated Regions and Their Roles in Viral Replication. Molecular Biology of the SARS-Coronavirus. 2009:47-61. doi:10.1007/978-3-642-03683-5_4.

20. Liu P, Millership JJ, Li L, Giedroc DP, Leibowitz JL. A Previously Unrecognized Unr Stem-Loop Structure in the Coronavirus 5' Untranslated Region Plays a Functional role in Replication. *Advances in Experimental Medicine and Biology The Nidoviruses*. 2006:25-30. doi:10.1007/978-0-387-33012-9_3.

21. Masters PS. The Molecular Biology of Coronaviruses. Advances in Virus Research. 2006:193-292. doi:10.1016/s0065-3527(06)66005-3.

22. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna*. 2004;10(8):1178-1190. doi:10.1261/rna.7650904.

23. Matsuyama S. Middle East Respiratory Syndrome (MERS) Coronavirus. *Journal of Veterinary Epidemiology*. 2013;17(2):112-116. doi:10.2743/jve.17.112.

24. Nourbakhsh M. Analysis of RNA Secondary Structure. *RNA Mapping Methods in Molecular Biology*. 2014:35-42. doi:10.1007/978-1-4939-1062-5_4.

25. Rice GM, Leonard CW, Weeks KM. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. Rna. 2014;20(6):846-854. doi:10.1261/rna.043323.113.

26. RNA Secondary Structure Prediction Including Pseudoknots, 2004; Lyngsø SpringerReference. doi:10.1007/springerreference_57865.

27. Sawicki SG. Coronavirus Genome Replication. *Viral Genome Replication*. 2009:25-39. doi:10.1007/b135974_2.

28. Sheahan TP, Baric RS. SARS Coronavirus Pathogenesis and Therapeutic Treatment Design. *Molecular Biology of the SARS-Coronavirus*. 2009:195-230. doi:10.1007/978-3-642-03683-5_13.

29. Shi ST, Lai MMC. Viral and Cellular Proteins Involved in Coronavirus Replication. *Current Topics in Microbiology and Immunology Coronavirus Replication and Reverse Genetics*.:95-131. doi:10.1007/3-540-26765-4_4.

30. Steffen I, Simmons G. Coronaviruses. eLS. 2015:1-9. doi:10.1002/9780470015902.a0023611.

31. Tahi F, Gouy M, Régnier M. Automatic RNA secondary structure prediction with a comparative approach. Computers & Chemistry. 2002;26(5):521-530. doi:10.1016/s0097-8485(02)00012-8.

32. Tian S, Cordero P, Kladwang W, Das R. Correcting a SHAPE-directed RNA structure by a mutate-map-rescue approach. 2014. doi:10.1101/001966.

33. Tong K-K, Cheung K-Y, Lee K-H, Leung K-S. Modified free energy model to improve RNA secondary structure prediction with pseudoknots. 13th IEEE International Conference on BioInformatics and BioEngineering. 2013. doi:10.1109/bibe.2013.6701532.

34. Ulasli M, Bayraktar R, Bozgeyik I. Replication of coronavirus. *Gaziantep Medical Journal*. 2013;19(3):141. doi:10.5455/gmj-30-2013-144.

35. Wbarthold S, Lsmith A. Mouse Hepatitis Virus. The Mouse in Biomedical Research. 2007:141-178. doi:10.1016/b978-012369454-6/50034-0.

36. Yount B, Denison MR, Weiss SR, Baric RS. Systematic Assembly of a Full-Length Infectious cDNA of Mouse Hepatitis Virus Strain A59. *Journal of Virology*. 2002;76(21):11065-11078. doi:10.1128/JVI.76.21.11065-11078.2002.

37. Zheng W-X, Chen L-L, Ou H-Y, Gao F, Zhang C-T. Coronavirus phylogeny based on a geometric approach. Molecular Phylogenetics and Evolution. 2005;36(2):224-232. doi:10.1016/j.ympev.2005.03.030.

APPENDIX

test_read_ct.R script

This script was used to extract selected areas of interest from the merged .ct file.

```
#ctfilename <- 'merged_BCoV.map_e1de.ct'
ctfilename <- 'merged_A591000.map_41eb.ct'

idxBegin <- 4600
idxEnd <- 5000

ctheader <- scan(ctfilename, what=list(numeric(0), character(0)), nlines=1,
quiet=TRUE)
ctoutputfilename <- paste0(ctfilename,"-cropped-from-",idxBegin,"-to-
",idxEnd,".ct")

ctdata <- read.table(ctfilename,header=FALSE, skip=1)
colnames(ctdata) <-
c("Index","Base","PrevIndex","NextIndex","PairedBase","Numbering")

ctdataCropped <- ctdata[idxBegin:idxEnd,]

idxOutOfRangePair <- (ctdataCropped$PairedBase<idxBegin &
ctdataCropped$PairedBase!=0) | ctdataCropped$PairedBase>idxEnd
ctdataCropped[which(idxOutOfRangePair),]$PairedBase <- 0

ctdataCropped$Index <- ctdataCropped$Index-idxBegin+1
ctdataCropped$PrevIndex <- ctdataCropped$PrevIndex-idxBegin+1
ctdataCropped$NextIndex <- ctdataCropped$NextIndex-idxBegin+1
idxPairedBases <- which(ctdataCropped$PairedBase!=0)
ctdataCropped[idxPairedBases,]$PairedBase <-
ctdataCropped[idxPairedBases,]$PairedBase-idxBegin+1

fileCon<-file(ctoutputfilename)
writeLines(paste(dim(ctdataCropped)[1],paste0(ctheader[[2]],"-cropped-from-
",idxBegin,"-to-",idxEnd)), fileCon)
close(fileCon)

write.table(ctdataCropped, append=TRUE,
file=ctoutputfilename,col.names=FALSE, row.names=FALSE, quote=FALSE)
```

Test_react_ct_batch.R script

This script was used to store the extracted sequence and connectivity data from the merged .ct file into separate .ct files, which were directly used in secondary structure visualization.

```
ctfilename <- 'merged_BCoV.map_e1de.ct'
```

Continued

```
loi_filename <-  'areas_of_interest_BCoV.csv'

#ctfilename <- 'merged_A591000.map_41eb.ct'
#loi_filename <- 'areas_of_interest_MHV.csv'



ctheader <- scan(ctfilename, what=list(numeric(0), character(0)), nlines=1,
quiet=TRUE)
ctdata <- read.table(ctfilename,header=FALSE, skip=1)
colnames(ctdata) <-
c("Index","Base","PrevIndex","NextIndex","PairedBase","Numbering")


locations_of_interest <- read.csv(loi_filename)


for (idx in 1:dim(locations_of_interest)[1]){

  idxBegin <- locations_of_interest$start[idx]
  idxEnd <- locations_of_interest$end[idx]

  ctoutputfilename <- paste0(ctfilename,"-cropped-from-",idxBegin,"-to-
",idxEnd,".ct")

  cat(idx,"out of",dim(locations_of_interest)[1],":
processing",ctoutputfilename,"...","\n")


  ctdataCropped <- ctdata[idxBegin:idxEnd,]

  idxOutOfRangePair <- (ctdataCropped$PairedBase<idxBegin &
ctdataCropped$PairedBase!=0) | ctdataCropped$PairedBase>idxEnd

  if(length(which(idxOutOfRangePair))>0){
    ctdataCropped[which(idxOutOfRangePair),]$PairedBase <- 0
  }

  ctdataCropped$Index <- ctdataCropped$Index-idxBegin+1
  ctdataCropped$PrevIndex <- ctdataCropped$PrevIndex-idxBegin+1
  ctdataCropped$NextIndex <- ctdataCropped$NextIndex-idxBegin+1
  idxPairedBases <- which(ctdataCropped$PairedBase!=0)
  ctdataCropped[idxPairedBases,]$PairedBase <-
ctdataCropped[idxPairedBases,]$PairedBase-idxBegin+1

  fileCon<-file(ctoutputfilename)
  writeLines(paste(dim(ctdataCropped)[1],paste0(ctheader[[2]],"-cropped-from-
",idxBegin,"-to-",idxEnd)), fileCon)
  close(fileCon)

  write.table(ctdataCropped, append=TRUE,
file=ctoutputfilename,col.names=FALSE, row.names=FALSE, quote=FALSE)

}
```